(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2011-221979
(P2011-221979A)
(43)公開日　平成23年11月4日(2011.11.4)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	350C	5B075
		G06F	17/30	210A	
		G06F	17/30	110C	

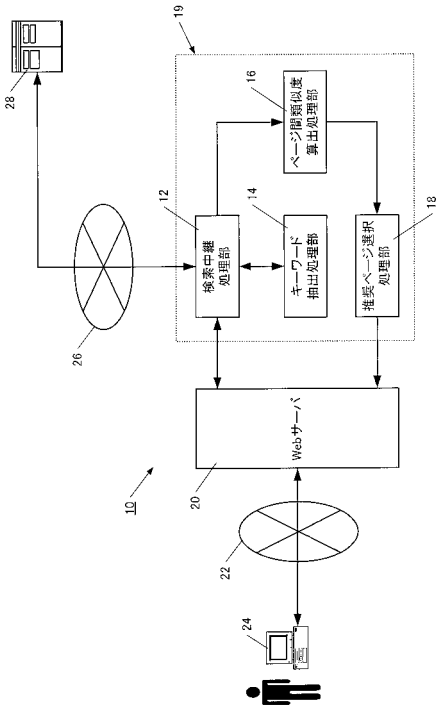
審査請求 未請求　請求項の数 5　ＯＬ　（全 14 頁）

(21)出願番号　特願2010-152556 (P2010-152556)
(22)出願日　平成22年7月3日 (2010.7.3)
(31)優先権主張番号　201010140447.0
(32)優先日　平成22年4月2日 (2010.4.2)
(33)優先権主張国　中国 (CN)

(71)出願人　502192546
清華大学
中華人民共和国北京市海淀区清華大学　郵編１００００８４
(71)出願人　000155469
株式会社野村総合研究所
東京都千代田区丸の内一丁目６番５号
(74)代理人　100096002
弁理士　奥田　弘之
(74)代理人　100091650
弁理士　奥田　規之
(72)発明者　李　春平
中華人民共和国北京市海淀区清華大学内
(72)発明者　王　益▲びん▼
中華人民共和国北京市海淀区清華大学内
最終頁に続く

(54)【発明の名称】検索支援システム、検索支援方法及び検索支援プログラム

(57)【要約】　（修正有）
【課題】ユーザの検索意図に合致したWebページを効率的に提示する。
【解決手段】検索キーワードを検索サーバ28に送信して検索を依頼する手段と、送信された検索結果リストをクライアント端末24に送信し、基準Webページの選択を促す手段と、この基準Webページ中のテキストを形態素単位に分解し、名詞をキーワードとして抽出する手段と、各抽出キーワードと検索キーワードを検索サーバ28に送信して検索を依頼する手段と、検索結果リストの中から、それぞれページランク順に上位２０件のWebページを類似候補ページとして抽出する手段と、各検索式に係る類似候補ページ間における類似度を算出する手段と、類似度の高い方から２０件の検索式に係る抽出キーワードを重要語と認定する手段と、全類似候補ページ中で重要語を３種以上含むものを類似ページと認定する手段と、類似ページリスト画面をクライアント端末24に送信する手段を備える。
【選択図】図１



【特許請求の範囲】
【請求項１】
クライアント端末から送信された検索キーワードを検索サーバに送信し、検索を依頼する手段と、
上記検索サーバから送信された検索結果リストが記載された画面を上記クライアント端末に送信し、基準Webページの選択を促す手段と、
上記クライアント端末から基準Webページの選択情報を受信した場合に、この基準Webページ中のテキストを形態素単位に分解し、特定の品詞に係るキーワードを抽出する手段と、
これらの抽出キーワードと上記検索キーワードとをAND条件で繋いだ検索式を抽出キーワード毎に生成し、これらの検索式を上記検索サーバに送信して検索を依頼する手段と、
検索サーバから送信された各検索式に係る検索結果リストの中から、それぞれページランク順に上位所定件数のWebページを類似候補ページとして抽出する手段と、
各検索式に係る類似候補ページ間における類似度を算出するページ間類似度算出手段と、
類似度の高い所定件数の検索式に係る抽出キーワードを、重要語と認定する手段と、
上記の全類似候補ページの中で、上記重要語を所定種類以上含むものを類似ページと認定する手段と、
この類似ページのリストが記載された画面を生成し、上記クライアント端末に送信する手段と、
を備えたことを特徴とする検索支援システム。
【請求項２】
上記ページ間類似度算出手段は、
各検索式に係る類似候補ページ中の一の類似候補ページを比較対象ページとして設定する処理と、この比較対象ページと残りの類似候補ページ間の類似度を個別に算出する処理と、各算出結果の中で類似度が上位所定数のものを抽出し、これらの類似度の平均値を隣接値として算出する処理を、当該検索式に係る全ての類似候補ページが比較対象ページとして設定されるまで繰り返した後、
得られた隣接値の中で最も大きな値の隣接値を当該検索式に係るページ間類似度と認定することを特徴とする請求項１に記載の検索支援システム。
【請求項３】
上記ページ間類似度算出手段は、
上記の比較対象ページと、他の類似候補ページを形態素単位に分解し、各ページから所定の品詞に係る形態素を抽出する処理と、抽出された各形態素のTF-IDF値を算出する処理と、この各形態素のTF-IDF値に基づいて各ページをベクトル化する処理と、比較対象ページのベクトルと他の類似候補ページのベクトルとの内積値を、両ページ間の類似度として算出する処理を実行することを特徴とする請求項２に記載の検索支援システム。
【請求項４】
クライアント端末から送信された検索キーワードを検索サーバに送信し、検索を依頼するステップと、
上記検索サーバから送信された検索結果リストが記載された画面を上記クライアント端末に送信し、基準Webページの選択を促すステップと、
上記クライアント端末から基準Webページの選択情報を受信した場合に、この基準Webページ中のテキストを形態素単位に分解し、特定の品詞に係るキーワードを抽出するステップと、
これらの抽出キーワードと上記検索キーワードとをAND条件で繋いだ検索式を抽出キーワード毎に生成し、これらの検索式を上記検索サーバに送信して検索を依頼するステップと、
検索サーバから送信された各検索式に係る検索結果リストの中から、それぞれページランク順に上位所定件数のWebページを類似候補ページとして抽出するステップと、

各検索式に係る類似候補ページ間における類似度を算出するページ間類似度算出ステップと、
類似度の高い所定件数の検索式に係る抽出キーワードを、重要語と認定するステップと、
上記の全類似候補ページの中で、上記重要語を所定種類以上含むものを類似ページと認定するステップと、
この類似ページのリストが記載された画面を生成し、上記クライアント端末に送信するステップと、
からなることを特徴とする検索支援方法。
【請求項５】
コンピュータを、
クライアント端末から送信された検索キーワードを検索サーバに送信し、検索を依頼する手段、
上記検索サーバから送信された検索結果リストが記載された画面を上記クライアント端末に送信し、基準Webページの選択を促す手段、
上記クライアント端末から基準Webページの選択情報を受信した場合に、この基準Webページ中のテキストを形態素単位に分解し、特定の品詞に係るキーワードを抽出する手段、
これらの抽出キーワードと上記検索キーワードとをAND条件で繋いだ検索式を抽出キーワード毎に生成し、これらの検索式を上記検索サーバに送信して検索を依頼する手段、
検索サーバから送信された各検索式に係る検索結果リストの中から、それぞれページランク順に上位所定件数のWebページを類似候補ページとして抽出する手段、
各検索式に係る類似候補ページ間における類似度を算出するページ間類似度算出手段、
類似度の高い所定件数の検索式に係る抽出キーワードを、重要語と認定する手段、
上記の全類似候補ページの中で、上記重要語を所定種類以上含むものを類似ページと認定する手段、
この類似ページのリストが記載された画面を生成し、上記クライアント端末に送信する手段、
として機能させることを特徴とする検索支援プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は検索支援システム、検索支援方法及び検索支援プログラムに係り、特に、ユーザの意図に合致したWebページを効率的に抽出する技術に関する。
【背景技術】
【０００２】
今日、Google（登録商標）やYahoo!（登録商標）といったインターネット上の検索サイトに検索キーワードを送信することにより、誰でもが手軽に様々な情報を入手することができる。
例えば、ある銘柄の株式の購入を検討している一般投資家が、当該銘柄の企業名を検索サイトの検索窓に入力して検索をリクエストすると、当該企業に関するWebページのリストが検索結果として送信され、Webブラウザ上に表示される。
これに対し投資家は、ニュース記事やリリース記事、ブログ記事等を次々と閲覧し、当該企業の最近の動向をチェックする。
そして、新製品情報や不祥事情報など、株価に影響を及ぼすような記事を見出した場合には、当該記事中に用いられている適当なキーワードを企業名に追加し、さらに検索を続けることで、必要な情報を集中的に収集することが可能となる。
【非特許文献１】Google　インターネットURL:http://www.google.co.jp/　検索日：平成２２年１月１６日
【非特許文献２】Yahoo! JAPAN　インターネットURL:http://www.yahoo.co.jp/　検索日：平成２２年１月１６日

【発明の開示】
【発明が解決しようとする課題】
【０００３】
しかしながら、このような従来の検索サイトを用いた検索方式の場合、ユーザには追加のキーワードを自分で選定して検索サイトに再投入し、イメージ通りの検索結果が得られるまで何度も操作を繰り返す手間がかかっていた。
この発明は、従来のこのような問題を解決するために案出されたものであり、ユーザの検索意図に合致したWebページを効率的に提示することができる技術の実現を目的としている。
【課題を解決するための手段】
【０００４】
上記の目的を達成するため、請求項１に記載した検索支援システムは、クライアント端末から送信された検索キーワードを検索サーバに送信し、検索を依頼する手段と、上記検索サーバから送信された検索結果リストが記載された画面を上記クライアント端末に送信し、基準Webページの選択を促す手段と、上記クライアント端末から基準Webページの選択情報を受信した場合に、この基準Webページ中のテキストを形態素単位に分解し、特定の品詞に係るキーワードを抽出する手段と、これらの抽出キーワードと上記検索キーワードとをAND条件で繋いだ検索式を抽出キーワード毎に生成し、これらの検索式を上記検索サーバに送信して検索を依頼する手段と、検索サーバから送信された各検索式に係る検索結果リストの中から、それぞれページランク順に上位所定件数のWebページを類似候補ページとして抽出する手段と、各検索式に係る類似候補ページ間における類似度を算出するページ間類似度算出手段と、類似度の高い所定件数の検索式に係る抽出キーワードを、重要語と認定する手段と、上記の全類似候補ページの中で、上記重要語を所定種類以上含むものを類似ページと認定する手段と、この類似ページのリストが記載された画面を生成し、上記クライアント端末に送信する手段とを備えたことを特徴としている。
【０００５】
請求項２に記載した検索支援システムは、請求項１のシステムを前提とし、さらに上記ページ間類似度算出手段が、各検索式に係る類似候補ページ中の一の類似候補ページを比較対象ページとして設定する処理と、この比較対象ページと残りの類似候補ページ間の類似度を個別に算出する処理と、各算出結果の中で類似度が上位所定数のものを抽出し、これらの類似度の平均値を隣接値として算出する処理を、当該検索式に係る全ての類似候補ページが比較対象ページとして設定されるまで繰り返した後、得られた隣接値の中で最も大きな値の隣接値を当該検索式に係るページ間類似度と認定することを特徴としている。
【０００６】
請求項３に記載した検索支援システムは、請求項２のシステムを前提とし、さらに上記ページ間類似度算出手段が、上記の比較対象ページと、他の類似候補ページを形態素単位に分解し、各ページから所定の品詞に係る形態素を抽出する処理と、抽出された各形態素のTF-IDF値を算出する処理と、この各形態素のTF-IDF値に基づいて各ページをベクトル化する処理と、比較対象ページのベクトルと他の類似候補ページのベクトルとの内積値を、両ページ間の類似度として算出する処理を実行することを特徴としている。
【０００７】
請求項４に記載した検索支援方法は、クライアント端末から送信された検索キーワードを検索サーバに送信し、検索を依頼するステップと、上記検索サーバから送信された検索結果リストが記載された画面を上記クライアント端末に送信し、基準Webページの選択を促すステップと、上記クライアント端末から基準Webページの選択情報を受信した場合に、この基準Webページ中のテキストを形態素単位に分解し、特定の品詞に係るキーワードを抽出するステップと、これらの抽出キーワードと上記検索キーワードとをAND条件で繋いだ検索式を抽出キーワード毎に生成し、これらの検索式を上記検索サーバに送信して検索を依頼するステップと、検索サーバから送信された各検索式に係る検索結果リストの中から、それぞれページランク順に上位所定件数のWebページを類似候補ページとして抽出

するステップと、各検索式に係る類似候補ページ間における類似度を算出するページ間類似度算出ステップと、類似度の高い所定件数の検索式に係る抽出キーワードを、重要語と認定するステップと、上記の全類似候補ページの中で、上記重要語を所定種類以上含むものを類似ページと認定するステップと、この類似ページのリストが記載された画面を生成し、上記クライアント端末に送信するステップとからなることを特徴としている。

【０００８】

請求項５に記載した検索支援プログラムは、コンピュータを、クライアント端末から送信された検索キーワードを検索サーバに送信し、検索を依頼する手段、上記検索サーバから送信された検索結果リストが記載された画面を上記クライアント端末に送信し、基準Webページの選択を促す手段、上記クライアント端末から基準Webページの選択情報を受信した場合に、この基準Webページ中のテキストを形態素単位に分解し、特定の品詞に係るキーワードを抽出する手段、これらの抽出キーワードと上記検索キーワードとをAND条件で繋いだ検索式を抽出キーワード毎に生成し、これらの検索式を上記検索サーバに送信して検索を依頼する手段、検索サーバから送信された各検索式に係る検索結果リストの中から、それぞれページランク順に上位所定件数のWebページを類似候補ページとして抽出する手段、各検索式に係る類似候補ページ間における類似度を算出するページ間類似度算出手段、類似度の高い所定件数の検索式に係る抽出キーワードを、重要語と認定する手段、上記の全類似候補ページの中で、上記重要語を所定種類以上含むものを類似ページと認定する手段、この類似ページのリストが記載された画面を生成し、上記クライアント端末に送信する手段として機能させることを特徴としている。

【発明の効果】

【０００９】

請求項１に記載の検索支援システム、請求項４に記載の検索支援方法及び請求項５に記載の検索支援プログラムによれば、ユーザは検索結果リスト中から自己の検索目的に最も近い一のWebページを基準Webページとして選択するだけで、当該Webページに類似したWebページのリストを取得することができるため、キーワードの選定及び再検索リクエストの煩わしさからユーザを解放することが可能となる。

【００１０】

請求項２に記載の検索支援システムの場合、検索キーワードと各抽出キーワードとの組合せ単位のページ間類似度を算出するに際し、まず一対の類似候補ページ同士で類似度を次々と算出していき、その過程で極端に低い値の類似度を排除した上で平均値を求め、各平均値の中で最大のものをページ間類似度として採用する方式を採用しているため、内容の空疎なノイズ的なページの影響を廃して、より信頼性の高いページ間類似度を導くことが可能となる。

【００１１】

請求項３に記載の検索支援システムの場合、両Webページに含まれる用語の構成や出現頻度に基づいて具体的な類似度が算出される仕組みであるため、記述内容に基づく類似性を算出結果に正確に反映させることが可能となる。

【発明を実施するための最良の形態】

【００１２】

図１は、この発明に係る検索支援システム10の機能構成を示すブロック図であり、検索中継処理部12と、キーワード抽出処理部14と、ページ間類似度算出処理部16と、推奨ページ選択処理部18とを備えた検索支援サーバ19と、Webサーバ20とから構成される。

上記の検索中継処理部12、キーワード抽出処理部14、ページ間類似度算出処理部16、推奨ページ選択処理部18は、検索支援サーバ19のCPUが、OS及びアプリケーションプログラムに従って必要な処理を実行することによって実現される。

【００１３】

検索中継処理部12及び推奨ページ選択処理部18は、Webサーバ20及びインターネット22を介して、ユーザの操作するクライアント端末24と接続される。

また、検索中継処理部12は、インターネット26を介して、Google（登録商標）やYahoo!

（登録商標）等の検索サーバ28と接続されている。この検索サーバ28には、インデックス情報を参照してユーザが入力した検索キーワードを含むWebページを抽出すると共に、抽出したWebページを所定のアルゴリズムに従ってランク付け（順位付け）する機能を備えた検索エンジンが搭載されている。

【００１４】

つぎに、図２のフローチャートに従い、この検索支援システム10における全体的な処理手順を説明する。

まず、一般ユーザがクライアント端末24を操作してWebサーバ20内の検索支援サイトにアクセスし、Webブラウザ上に表示された検索窓に検索キーワードを入力して検索ボタンをクリックすると、クライアント端末24からWebサーバ20に検索キーワードが送信される。

【００１５】

この検索キーワードをWebサーバ20経由で受信した検索中継処理部12は（Ｓ10）、これを検索サーバ28に送信し、検索を依頼する（Ｓ12）。

そして、検索サーバ28から検索結果リストが送信されると、これを受けた検索中継処理部12は（Ｓ14）、検索結果リストをWebサーバ20に送信する。

Webサーバ20は、この検索結果リストが記述された画面（Htmlファイル）を生成し、クライアント端末24に送信する（Ｓ16）。

【００１６】

この結果、図３に示すように、検索結果リスト画面30がクライアント端末24のWebブラウザ上に表示される。

この図においては、「野村総合研究所」という検索キーワードの投入に対して、「野村総合研究所（NRI）」、「採用情報」、「NRI-Wikipedia」等のタイトルと、それぞれの概要情報が、検索結果リストとして列挙されている。

【００１７】

これに対しユーザは、各タイトルをクリックして対応のWebページをWebブラウザ上に表示させ、個別に内容をチェックする。

そして、注目すべき内容を備えたWebページを発見した場合、上記の検索結果リスト画面30に戻り、当該Webページのタイトル近傍に設けられたチェックボックス32にチェックを入れた上で、「類似ページを見る」ボタン34をクリックする。

この結果、クライアント端末24からWebサーバ20に対して、基準Webページの選択情報を伴う類似ページの検索リクエストが送信される。

【００１８】

Webサーバ20からこの基準Webページの選択情報を受け取った検索中継処理部12は（Ｓ18）、基準WebページのURLにアクセスして当該Webページを取得し（Ｓ19）、キーワード抽出処理部14に渡す。

これを受けたキーワード抽出処理部14は、当該基準ページに含まれるテキストに対して形態素解析処理を実行する（Ｓ20）。

ここで「形態素解析処理」とは、自然言語で記述された文を、意味を有する最小の言語単位である形態素に分解し、それぞれの品詞を特定する処理をいう。

【００１９】

つぎにキーワード抽出処理部14は、各形態素の中から特定の品詞に係るものを抽出する（Ｓ22）。ここでは、名詞の形態素が９０個抽出されたものとして、話を進める。

【００２０】

つぎに、キーワード抽出処理部14から抽出キーワードを渡された検索中継処理部12は、ユーザが最初に入力した検索キーワードと、各抽出キーワードとをAND条件で繋いだ９０個分の検索式を生成し、各検索式を検索サーバ28に送信して検索を依頼する（Ｓ24）。

図４は、この場合の具体例を示すものであり、当初の検索キーワードである「野村総合研究所」に対して、「本日」、「新サービス」、「発表」等の抽出キーワードがスペースを介してAND条件で結ばれた状態で、検索サーバ28に投入されるイメージを表している。

【００２１】
そして、検索サーバ28から「検索キーワード＆抽出キーワード」単位での検索結果を受信すると（Ｓ26）、検索中継処理部12はページランクが上位２０件以内のWebページを類似候補ページとして抽出し（Ｓ28）、ページ間類似度算出処理部16に渡す。
図４には、９０組の検索式（検索キーワード＆抽出キーワード）についてそれぞれ２０件のWebページが抽出された結果、20（件）×90（組）＝1,800（ページ）の類似候補ページ群が得られた例が示されている。
【００２２】
上記のページランクは、検索結果である各Webページの被リンク数やリンク元ページのページランク等に基づき、固有のアルゴリズムに従って検索サーバ28が付与するものであり、各Webページの有用性を示す指標と評価できる。
このため、このシステム10においては、ページランクの上位２０件を類似候補ページとして抽出することにより、有用性の低いWebページを排除している。ただし、この抽出件数の「２０件」は一例であり、他の閾値（件数）を適用することも当然に可能である。また、固定的な件数の代わりに、所定のページランク以上のWebページを全て類似候補ページとして抽出することもできる。
【００２３】
検索中継処理部12から類似候補ページ群を渡されたページ間類似度算出処理部16は、検索式（検索キーワード＆抽出キーワード）単位で、上位２０件の類似候補ページ間における類似度を算出する（Ｓ30）。このページ間類似度の具体的な算出手順については、後に詳述する。
【００２４】
ページ間類似度算出処理部16から、合計９０組分のページ間類似度の算出結果を受け取った推奨ページ選択処理部18は、各ページ間類似度の数値を高い順に整列させ、上位２０組の検索式に係る抽出キーワードを「重要語」と認定する（Ｓ32）。図４においては、９０組の検索式から、「新サービス」、「アメリカ企業」、「コスト削減」等が重要語として抽出された例が示されている。
なお、この「上位２０組」も一例であり、他の件数の抽出キーワードを重要語として抽出してもよい。
【００２５】
つぎに推奨ページ選択処理部18は、1,800ページに及ぶ類似候補ページ群の中から、上記２０個の重要語の中で３種以上の重要語を含むものを類似ページとして選定し、類似ページリストを生成する（Ｓ34）。
各重要語は、当初の検索キーワード（野村総合研究所）との組合せにおいて、基準Webページの内容を特徴付ける文字通り重要な意義を有する語であると考えられるため、より多くの種類の重要語を含むWebページは、それだけ基準Webページに対する類似度が高いものと評価できる。
ただし、「３種以上」の閾値は一例であり、他の値を閾値として設定することができる。また、複数種類の重要語を含むことが必要とされるため、一種類の重要語（例えば「コスト削減」）のみが１００回登場するようなWebページは、類似ページと認定されることはない。
【００２６】
推奨ページ選択処理部18からこの類似ページリストを受け取ったWebサーバ20は、この類似ページリストを含む画面（Htmlファイル）を生成し、クライアント端末24に送信する（Ｓ36）。
【００２７】
この結果、図５に示すように、クライアント端末24のWebブラウザ上に類似ページリスト画面40が表示される。
この類似ページリスト中のタイトルをユーザがクリックすると、ユーザが最初に選択した基準Webページの記述内容と類似した内容を備えたWebページが表示されることとなる。

例えば、発明者が行った実証試験の結果では、適合率（基準Webページと関係のあるページが含まれる割合）が平均で８６％、再現率（重要なページが含まれる割合）が平均で６０％となり、かなり高い精度であることが確認された。

【００２８】

上記においては、一のWebページが基準Webページとして選択された例を示したが、ユーザは、図３の検索結果リスト画面30において２以上のチェックボックス32にチェックを入れることにより、複数の基準Webページを選択することもできる。

この場合、キーワード抽出処理部14は各基準Webページを形態素に分解し、特定品詞（例えば名詞）に係るキーワードを各基準Webページから抽出した上で、これらを一括の抽出キーワード群となす。

後は、検索中継処理部12等により、図２のＳ24以下の処理が順次実行されることにより、類似ページリスト画面40が生成され、クライアント端末24に送信される。

【００２９】

つぎに、図６のフローチャート及び図７の説明図に従い、ページ間類似度算出処理部16によるページ間類似度の算出手順について説明する。

まずページ間類似度算出処理部16は、２０件の類似候補ページの中から一のページを、比較対象ページとして設定する（Ｓ30-01）。図７(a)では、ページＡが最初の比較対象ページに設定された例が示されている。

【００３０】

つぎにページ間類似度算出処理部16は、TF-IDF及びベクトル空間法を用いて、残り１９件のページ（ページＢ～ページＴ）とページＡとの間の類似度を個別に算出する（Ｓ30-02）。このTF-IDF及びベクトル空間法を用いた類似度の具体的な算出手順については、後に詳述する。

【００３１】

つぎにページ間類似度算出処理部16は、算出された計１９個の類似度を高い順に整列させ、上位１５個の類似度の平均値を隣接値として算出する（Ｓ30-03）。図７(a)においては、ページＡを比較対象とした場合の隣接値として「0.21」の数値が導かれている。

【００３２】

つぎにページ間類似度算出処理部16は、図７(b)に示すように、次のページＢを比較対象に設定した上で（Ｓ30-04／Ｎ、Ｓ30-01）、上記と同様の手順に従い、隣接値を求める（Ｓ30-02、Ｓ30-03）。

そして、図７(c)に示すように、最後のページＴを比較対象に設定した上で、上記と同様の手順に従って隣接値を求め終えた時点で（Ｓ30-04／Ｙ）、ページ間類似度算出処理部16は全２０件の隣接値の中で最大の値を備えたものを、当該検索式に係るページ間類似度と認定する（Ｓ30-05）。

図７においては、ある検索式に係るページ間類似度として、「0.32」が導かれた例が示されている。

【００３３】

つぎに、図８のフローチャート及び図９、図１０の説明図に従い、TF-IDF及びベクトル空間法を用いた類似度の具体的な算出手順について説明する。

まずページ間類似度算出処理部16は、各類似候補ページに対して形態素解析を行い、特定品詞（例えば名詞）の用語のみを抽出する（Ｓ30-02-01）。

図９の例では、ページαの「今日が締め切りだ。今日も徹夜かな。」から「今日／締め切り／今日／徹夜」の用語が、ページβの「今日も煮干しだ。飽き飽きだ。」から「今日／煮干し」の用語が、ページγの「今日は天気がよい。野球をしよう。」から「今日／天気／野球」の用語が、ページδの「天気がよい。サッカーをしよう。」から「天気／サッカー」の用語がそれぞれ取り出されている。なお、図９及び図１０においては、説明を単純化するためページα～ページδの４件の類似候補ページに基づいてページ間類似度を算出する例が示されているが、類似候補ページの数が増えた場合も基本的な考え方は同じである。

【００３４】
つぎにページ間類似度算出処理部16は、各ページにおける各用語の頻度（TF／Term Frequency）を算出する（Ｓ30-02-02）。例えば、ページαにおける「今日」の頻度は「２」となる。
【００３５】
つぎにページ間類似度算出処理部16は、用語毎に当該用語を含むページ数（DF／Document Frequency）を算出し（Ｓ30-02-03）、DF辞書50に格納する（Ｓ30-02-04）。例えば、ページα～δにおいて「今日」を含むページは３件あるため、「今日」のDFは「３」となる。
【００３６】
つぎにページ間類似度算出処理部16は、このDF辞書50に基づいて各ページをベクトル化する。
例えば、ページαの場合はDF辞書50に収録された用語の中、「今日」「締め切り」「徹夜」の３種類の用語を含んでいるため、ページ間類似度算出処理部16はこれらの用語のDFに基づいて、IDF（Inverse Document Frequency）及びTF-IDFを求める。
【００３７】
まずページ間類似度算出処理部16は、以下のようにして各用語のIDFを算出する（Ｓ30-02-05）。

IDF（今日）＝log（文書数／DF）
＝log（４／３）

【００３８】
つぎにページ間類似度算出処理部16は、以下のようにして各用語のTF-IDFを算出する（Ｓ30-02-06）。

TF-IDF（今日）＝TF（今日）×IDF（今日）
＝２×log（４／３）＝0.25

同様の処理により、ページ間類似度算出処理部16は「締め切り」のTF-IDF＝0.60、「徹夜」のTF-IDF＝0.60を算出する。
【００３９】
ここで、ページαに含まれる「今日」「締め切り」「徹夜」の３種類の用語はDF辞書50における掲載順が１～３番であるため、図１０に示すように、ベクトル要素として１～３行までに0.25、0.60、0.60の数値が代入され、他の用語の掲載順に対応する行には0.00が代入されたベクトルがページ間類似度算出処理部16によって生成され、ページαのベクトルとなされる（Ｓ30-02-07）。
【００４０】
このページαのベクトル長は0.89であるため、ページ間類似度算出処理部16はベクトル長を１に揃えるための正規化処理を各数値に対して施し（Ｓ30-02-08）、最終的に0.28、0.68、0.68、0.00、0.00、0.00、0.00の数値が充填されたベクトルを導く。
【００４１】
ページβの場合にはDF辞書50に収録された用語の中、「今日」「煮干し」の２種類の用語を含んでおり、これらの用語のDF辞書50における掲載順が１番と４番であるため、ベクトル要素として１行目及び４行目に0.12及び0.60の数値が代入され、他の用語の掲載順に対応する行には0.00が代入されたベクトルが生成された後（Ｓ30-02-07）、上記と同様の正規化処理により（Ｓ30-02-08）、最終的に0.20、0.00、0.00、0.98、0.00、0.00、0.00の数値が充填されたベクトルが得られる。
【００４２】
また、文書γの場合はDF辞書50に収録された用語の中、「今日」「天気」「野球」の３種類の用語を含んでおり、これらの用語のDF辞書50における掲載順が１番と５番、６番であるため、ベクトル要素として１行目、５行目、６行目にそれぞれ0.12、0.60、0.30の数値が代入され、他の用語の掲載順に対応する行には0.00が代入されたベクトルが生成された後（Ｓ30-02-07）、上記と同様の正規化処理により（Ｓ30-02-08）、最終的に0.18、0.

00、0.00、0.00、0.88、0.44、0.00の数値が充填されたベクトルが得られる。

【００４３】

また、ページδの場合はDF辞書50に収録された用語の中、「天気」「サッカー」の２種類の用語を含んでおり、これらの用語のDF辞書50における掲載順が６番と７番であるため、ベクトル要素として６行目及び７行目にそれぞれ0.30、0.60の数値が代入され、他の用語の掲載順に対応する行には0.00が代入されたベクトルが生成された後（Ｓ30-02-07）、上記と同様の正規化処理により（Ｓ30-02-08）、最終的に0.00、0.00、0.00、0.00、0.00、0.45、0.89の数値が充填されたベクトルが得られる。

【００４４】

つぎにページ間類似度算出処理部16は、ページαのベクトルとページβのベクトルとの間の内積（距離）を求める（Ｓ30-02-09）。この内積値が、両ページ間の類似度となる。

以後、ページ間類似度算出処理部16は同様の手順に従い、ページα－ページγ間の類似度及びページα－ページδ間の類似度を算出する。

【図面の簡単な説明】

【００４５】

【図１】この発明に係る検索支援システムの機能構成を示すブロック図である。

【図２】この検索支援システムの全体的な処理手順を示すフローチャートである。

【図３】検索結果リスト画面を示す図である。

【図４】類似ページの抽出に係る手順を示す概念図である。

【図５】類似ページリスト画面を示す図である。

【図６】ページ間類似度の算出手順を示すフローチャートである。

【図７】ページ間類似度の算出手順を示す概念図である。

【図８】TF-IDF及びベクトル空間法を用いた類似度の具体的な算出手順を示すフローチャートである。

【図９】TF-IDF及びベクトル空間法を用いた類似度の具体的な算出手順を示す説明図である。

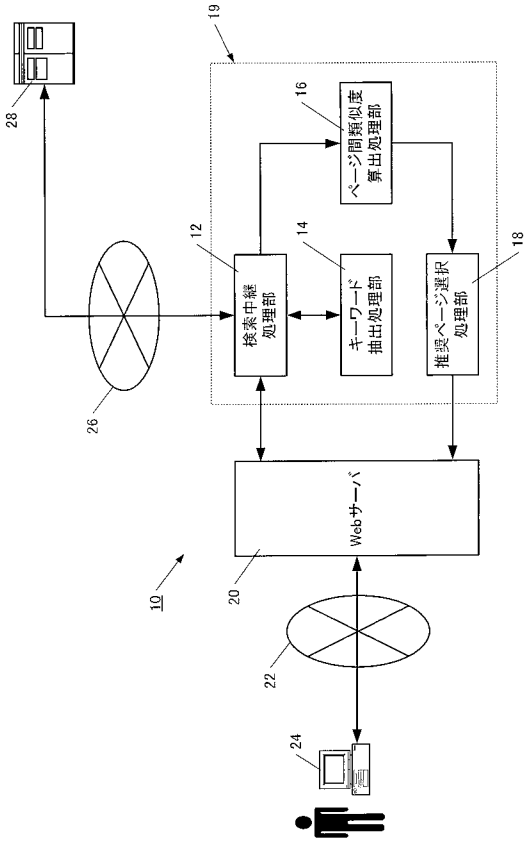
【図１０】TF-IDF及びベクトル空間法を用いた類似度の具体的な算出手順を示す説明図である。

【符号の説明】

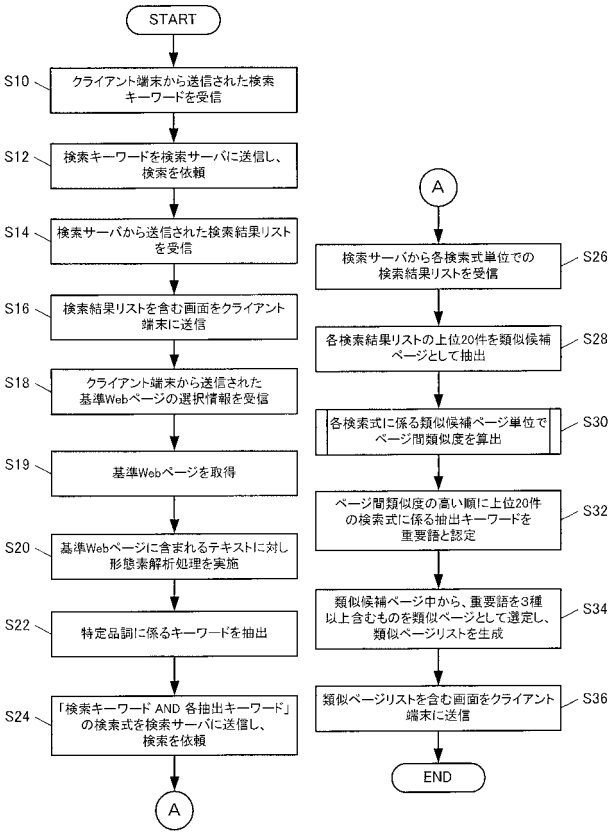
【００４６】

10　検索支援システム
12　検索中継処理部
14　キーワード抽出処理部
16　ページ間類似度算出処理部
18　推奨ページ選択処理部
19　検索支援サーバ
20　Webサーバ
22　インターネット
24　クライアント端末
26　インターネット
28　検索サーバ
30　検索結果リスト画面
32　チェックボックス
34　「類似ページを見る」ボタン
40　類似ページリスト画面
50　DF辞書

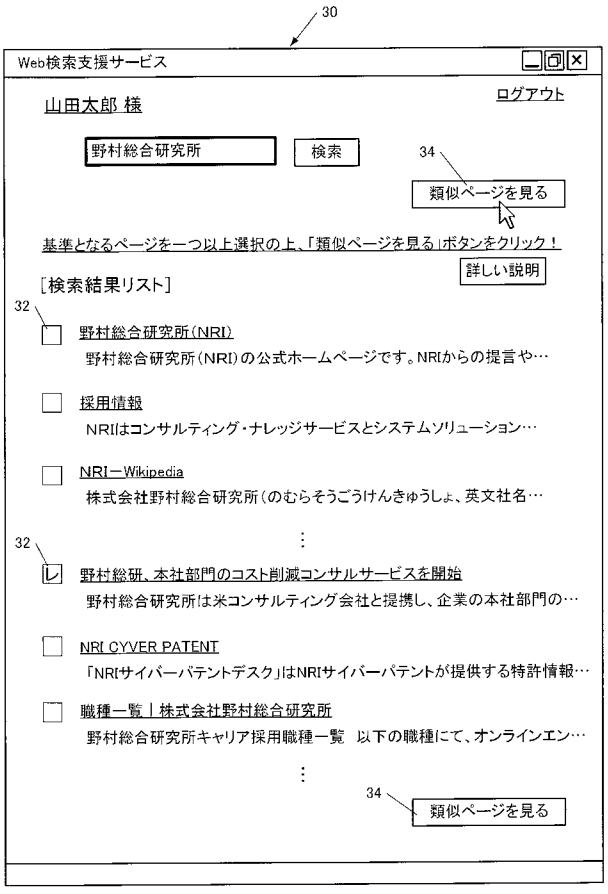
【図１】


19
28
16
ページ間類似度算出処理部
12
14
18
検索中継処理部
キーワード抽出処理部
推奨ページ選択処理部
26
Webサーバ
10
20
22
24

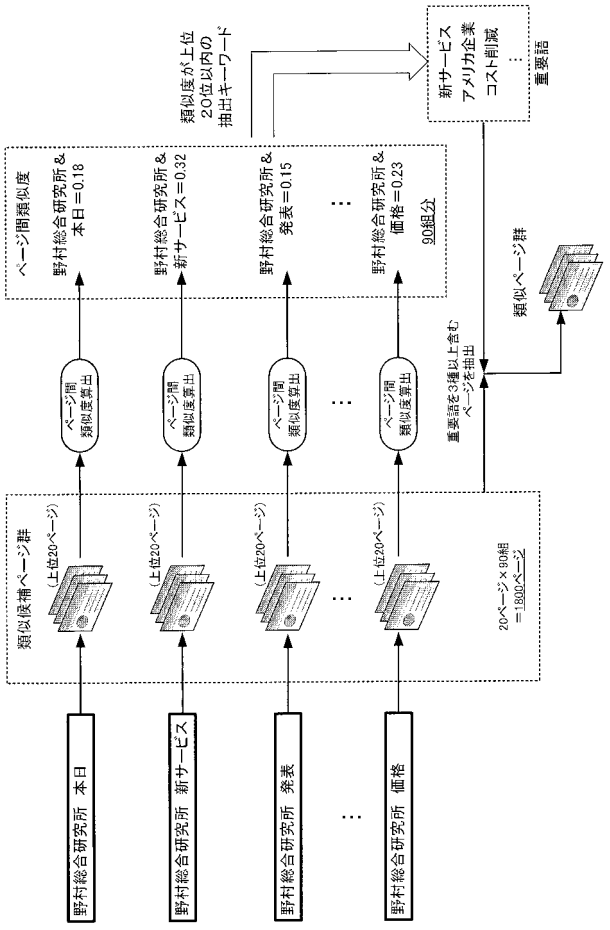
【図２】


START
S10
クライアント端末から送信された検索キーワードを受信
S12
検索キーワードを検索サーバに送信し、検索を依頼
S14
検索サーバから送信された検索結果リストを受信
S16
検索結果リストを含む画面をクライアント端末に送信
S18
クライアント端末から送信された基準Webページの選択情報を受信
S19
基準Webページを取得
S20
基準Webページに含まれるテキストに対し形態素解析処理を実施
S22
特定品詞に係るキーワードを抽出
S24
「検索キーワード AND 各抽出キーワード」の検索式を検索サーバに送信し、検索を依頼
A
A
S26
検索サーバから各検索式単位での検索結果リストを受信
S28
各検索結果リストの上位20件を類似候補ページとして抽出
S30
各検索式に係る類似候補ページ単位でページ間類似度を算出
S32
ページ間類似度の高い順に上位20件の検索式に係る抽出キーワードを重要語と認定
S34
類似候補ページ中から、重要語を3種以上含むものを類似ページとして選定し、類似ページリストを生成
S36
類似ページリストを含む画面をクライアント端末に送信
END

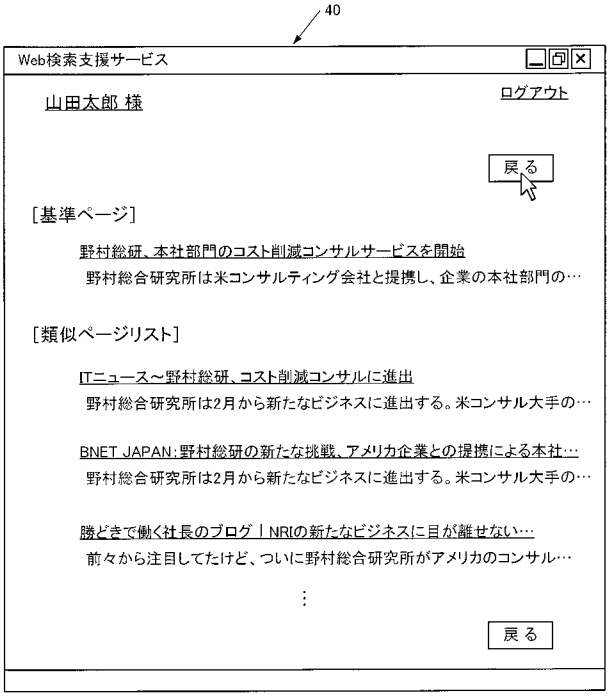
【図３】


30
Web検索支援サービス
山田太郎 様
ログアウト
野村総合研究所
検索
34
類似ページを見る
基準となるページを一つ以上選択の上、「類似ページを見る」ボタンをクリック！
詳しい説明
[検索結果リスト]
32
野村総合研究所(NRI)
野村総合研究所(NRI)の公式ホームページです。NRIからの提言や…
採用情報
NRIはコンサルティング・ナレッジサービスとシステムソリューション…
NRI－Wikipedia
株式会社野村総合研究所（のむらそうごうけんきゅうしょ、英文社名…
32
野村総研、本社部門のコスト削減コンサルサービスを開始
野村総合研究所は米コンサルティング会社と提携し、企業の本社部門の…
NRI CYVER PATENT
「NRIサイバーパテントデスク」はNRIサイバーパテントが提供する特許情報…
職種一覧 | 株式会社野村総合研究所
野村総合研究所キャリア採用職種一覧　以下の職種にて、オンラインエン…
34
類似ページを見る

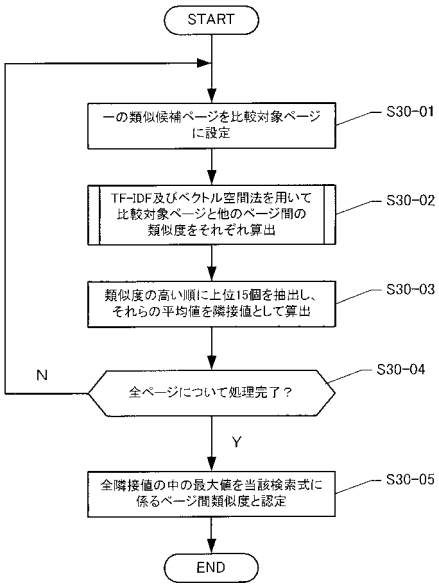
【図４】


類似度が上位20位以内の抽出キーワード
新サービス
アメリカ企業
コスト削減
…
重要語
ページ間類似度
野村総合研究所＆本日＝0.18
野村総合研究所＆新サービス＝0.32
野村総合研究所＆発表＝0.15
野村総合研究所＆価格＝0.23
90組分
類似ページ群
ページ間類似度算出
重要語を3種以上含むページを抽出
類似候補ページ群
(上位20ページ)
20ページ×90組＝1800ページ
野村総合研究所 本日
野村総合研究所 新サービス
野村総合研究所 発表
野村総合研究所 価格

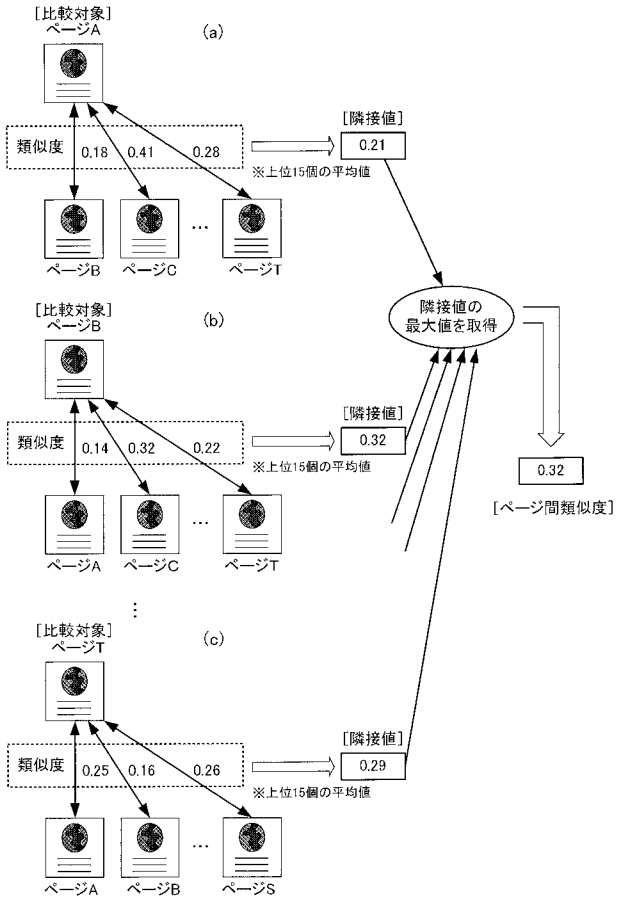
【図５】


40
Web検索支援サービス
山田太郎 様
ログアウト
戻る
[基準ページ]
野村総研、本社部門のコスト削減コンサルサービスを開始
野村総合研究所は米コンサルティング会社と提携し、企業の本社部門の…
[類似ページリスト]
ITニュース～野村総研、コスト削減コンサルに進出
野村総合研究所は2月から新たなビジネスに進出する。米コンサル大手の…
BNET JAPAN：野村総研の新たな挑戦、アメリカ企業との提携による本社…
野村総合研究所は2月から新たなビジネスに進出する。米コンサル大手の…
勝どきで働く社長のブログ | NRIの新たなビジネスに目が離せない…
前々から注目してたけど、ついに野村総合研究所がアメリカのコンサル…
戻る

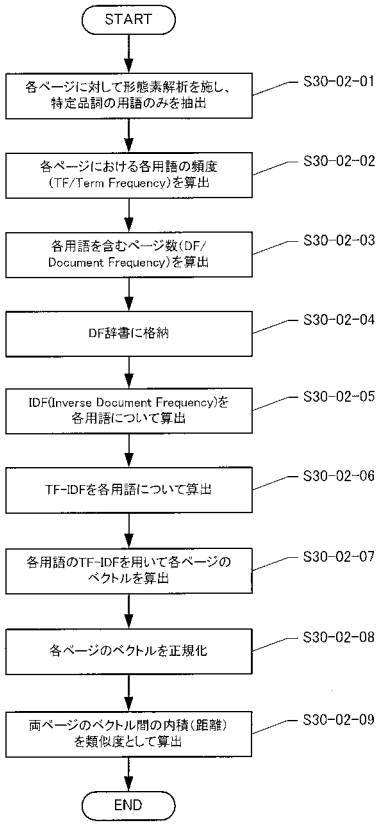
【図６】


START
一の類似候補ページを比較対象ページに設定
S30-01
TF-IDF及びベクトル空間法を用いて比較対象ページと他のページ間の類似度をそれぞれ算出
S30-02
類似度の高い順に上位15個を抽出し、それらの平均値を隣接値として算出
S30-03
全ページについて処理完了？
S30-04
N
Y
全隣接値の中の最大値を当該検索式に係るページ間類似度と認定
S30-05
END

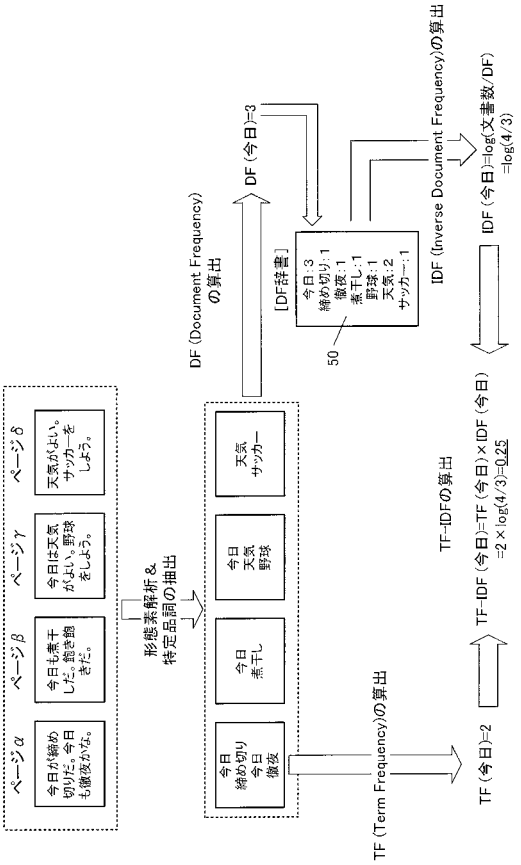
【図７】


[比較対象]
ページA
(a)
類似度
0.18
0.41
0.28
[隣接値]
0.21
※上位15個の平均値
ページB
ページC
ページT
[比較対象]
ページB
(b)
隣接値の最大値を取得
類似度
0.14
0.32
0.22
[隣接値]
0.32
※上位15個の平均値
0.32
[ページ間類似度]
ページA
ページC
ページT
[比較対象]
ページT
(c)
類似度
0.25
0.16
0.26
[隣接値]
0.29
※上位15個の平均値
ページA
ページB
ページS

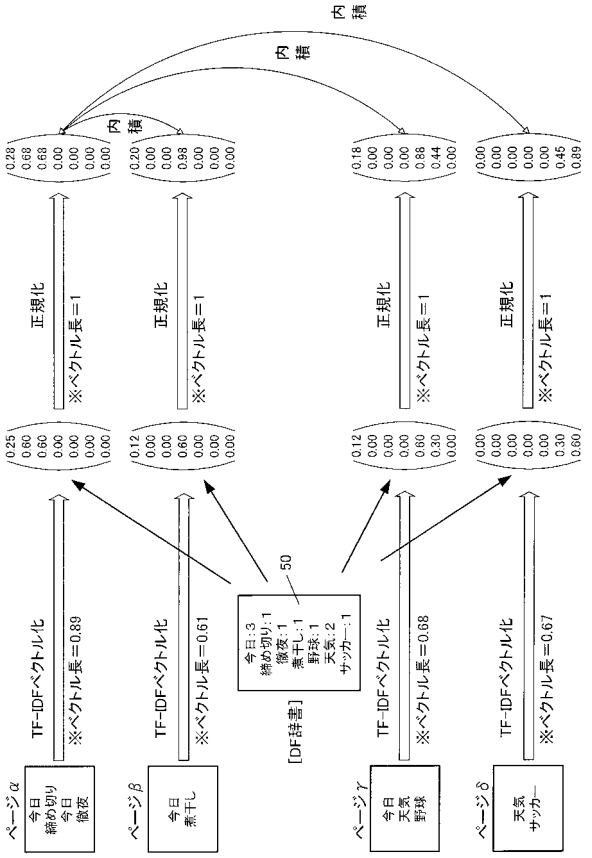
【図８】


START
各ページに対して形態素解析を施し、特定品詞の用語のみを抽出
S30-02-01
各ページにおける各用語の頻度(TF/Term Frequency)を算出
S30-02-02
各用語を含むページ数(DF/Document Frequency)を算出
S30-02-03
DF辞書に格納
S30-02-04
IDF(Inverse Document Frequency)を各用語について算出
S30-02-05
TF-IDFを各用語について算出
S30-02-06
各用語のTF-IDFを用いて各ページのベクトルを算出
S30-02-07
各ページのベクトルを正規化
S30-02-08
両ページのベクトル間の内積(距離)を類似度として算出
S30-02-09
END

【図９】


ページα
今日が締め切りだ。今日も徹夜かな。
ページβ
今日も煮干しだ。腐き飽きだ。
ページγ
今日は天気がよい。野球をしよう。
ページδ
天気がよい。サッカーをしよう。
形態素解析&
特定品詞の抽出
今日
締め切り
今日
徹夜
今日
煮干し
今日
天気
野球
天気
サッカー
DF (Document Frequency)
の算出
DF (今日)=3
[DF辞書]
今日:3
締め切り:1
徹夜:1
煮干し:1
野球:1
天気:2
サッカー:1
50
IDF (Inverse Document Frequency)の算出
IDF (今日)=log(文書数/DF)
=log(4/3)
TF-IDFの算出
TF (Term Frequency)の算出
TF (今日)=2
TF-IDF (今日)=TF (今日)×IDF (今日)
=2×log(4/3)=0.25

【図１０】


内積
内積
内積
正規化
※ベクトル長=1
TF-IDFベクトル化
※ベクトル長=0.89
※ベクトル長=0.61
※ベクトル長=0.68
※ベクトル長=0.67
50
[DF辞書]
今日:3
締め切り:1
徹夜:1
煮干し:1
野球:1
天気:2
サッカー:1
ページα
今日
締め切り
今日
徹夜
ページβ
今日
煮干し
ページγ
今日
天気
野球
ページδ
天気
サッカー

フロントページの続き

(72)発明者　阿部　昌平
東京都千代田区丸の内一丁目６番５号　株式会社野村総合研究所内

Ｆターム(参考) 5B075 KK02　ND03　NK32　QM05